W602–W605 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh375

iMOT: an interactive package for the selection of spatially interacting motifs

A. Bhaduri, G. Pugalenthi, N. Gupta[1] and R. Sowdhamini*

National Centre for Biological Sciences, Tata Institute of Fundamental Research, UAS-GKVK Campus, Bellary Road, Bangalore 560 065, Karnataka, India and [1]Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur 208016, Uttar Pradesh, India

Received February 14, 2004; Revised and Accepted March 15, 2004

ABSTRACT

Functional selection and three-dimensional structural constraints of proteins relate to the retention of significant sequence similarity between proteins of similar fold and function despite poor overall sequence identity and evolutionary pressures. We report the availability of 'iMOT' (interacting MOTif) server, an interactive package for the automatic identification of spatially interacting motifs among distantly related proteins sharing similar folds and possessing common ancestral lineage. Spatial interactions between conserved stretches of a protein are evaluated by calculations of pseudo-potentials that describe the strength of interactions. Such an evaluation permits the automatic identification of highly interacting conserved regions of a protein. Interacting motifs have been shown to be useful in searching for distant homologues and establishing remote homologies among the largely unassigned sequences in genome databases. Information on such motifs should also be of value in protein folding, modelling and engineering experiments. The iMOT server can be accessed from http://www.ncbs.res.in/~faculty/mini/imot/iMOTserver.html. Supplementary Material can be accessed from: http://www.ncbs.res.in/~faculty/mini/imot/supplementary.html.

INTRODUCTION

The final folded form of proteins is determined by the primary structure (1). The past decade has been marked by an exponential rise in sequence and structural information on proteins (2). Proteins belonging to homologous families are known to share similar descriptors or critical residues (3). Several proteins share insignificant sequential identity but are structurally similar (4) and possess evolutionary relationship (5,6).

PROSITE (7), BLOCKS (8) and PRINTS (9) are databases that record large collections of biologically meaningful signatures that are described as patterns but are usually restricted to sequentially contiguous, structural or functional residues conserved across members of well-characterized families. The analysis of structural conserved spheres around homologous protein structures constitutes the first attempt to predict structural motifs that are preserved among family members (10). Structural descriptors for proteins related at the superfamily level are hard to obtain.

Our approach to identifying conserved spatially interacting regions across proteins (11) is described briefly below. When demonstrating the applications of these signatures for establishing remote homology between distantly related proteins (11), such motifs were selected manually and are thus limited for large-scale analysis. In this paper, we report the further inclusion of pseudo-energies that permits the automatic identification of spatially interacting motifs. This is presented via a user-friendly web interface.

IDENTIFYING INTERACTING MOTIFS AMONG KNOWN STRUCTURES

In the automated approach to identifying spatially interacting motifs, the following three steps have been followed: (i) a search for homologues against sequence databases was performed using PSI-BLAST (12); (ii) conserved residues or motifs were recognized by examining the amino acid exchanges (13) within the respective homologues; (iii) pseudo-energies, which include components of electrostatic, steric and hydrophobic interactions, were calculated across conserved stretches of a protein (14).

Searching for homologues

Sequence homologues for each of the query sequences can be identified from the SWISS-PROT database (15), PDB sequence database (16) or the Non Redundant protein sequence database using BLAST (12). The hits were filtered for redundancies, and proteins no more than 60% identical were aligned with the query using MALIGN (17).

*To whom correspondence should be addressed. Tel: +91 80 3636421/8 Ext; Fax: +91 80 3636662; Email: mini@ncbs.res.in

Determining conservation among residues

The amino acid substitutions at each alignment position were scored by referring to a symmetric amino acid exchange matrix derived from several alignments of homologous protein structures (18–20). Regions that contain at least two conserved residues (with a score >50) in a window of five were considered further. For those queries where there are no sequence homologues (even after a further search in the non-redundant database), each residue position was weighed equally and was analysed for scoring the interactions.

Evaluation of the Interaction

Following Novotny and co-workers' approach (21,22), the interaction among the motifs is viewed to be a manifestation of three biophysical effects: the hydrophobic, the steric and the electrostatic. Electrostatic interaction is calculated using the Debye–Huckel expression, similar to the approach adapted by Crichton and co-workers (23). The hydrophobic (22) and the steric interactions (24) are evaluated based on previously published approaches by other groups.

THE iMOT SERVER

The iMOT server can be accessed from http://www.ncbs.res.in/~faculty/mini/imot/iMOTserver.html. In order to view the interacting motifs of the query protein, the user may submit the sequence of a given protein in the PIR format. Options are provided to obtain homologues for the query from various sequence databases (the PDB sequence database, the SWISS-PROT database and the Non Redundant protein sequence database) using the BLAST program at different levels of stringency through expectation values (Figure 1).

Projection of results

The critical residues contained in the interacting motifs are tabulated and displayed for each of the query sequences. The position of the motifs in the provided sequence and the actual residue number as mentioned in the PDB are also listed. Common interacting motifs for the sequence alignment of structures are tabulated for viewing the common fingerprints for the set of query proteins. Options for viewing the motifs on the alignment submitted are also provided. Although iMOT is designed for multiple structural members, the user can provide the query sequence alone and the nearest structural homologues can be identified internally by running BLAST (12) against PDB (16) and aligned. The interacting motifs, as explained above, are identified for the structural homologue to report the descriptors conserved in the sequence (Figure 2). In cases where only sequence homologues are present, the sequentially conserved regions alone (identified by same method as mentioned above) are reported.

Interacting motifs and PROSITE database

To make the comparison with PROSITE motifs, we considered a non-redundant database of 3867 structural domains

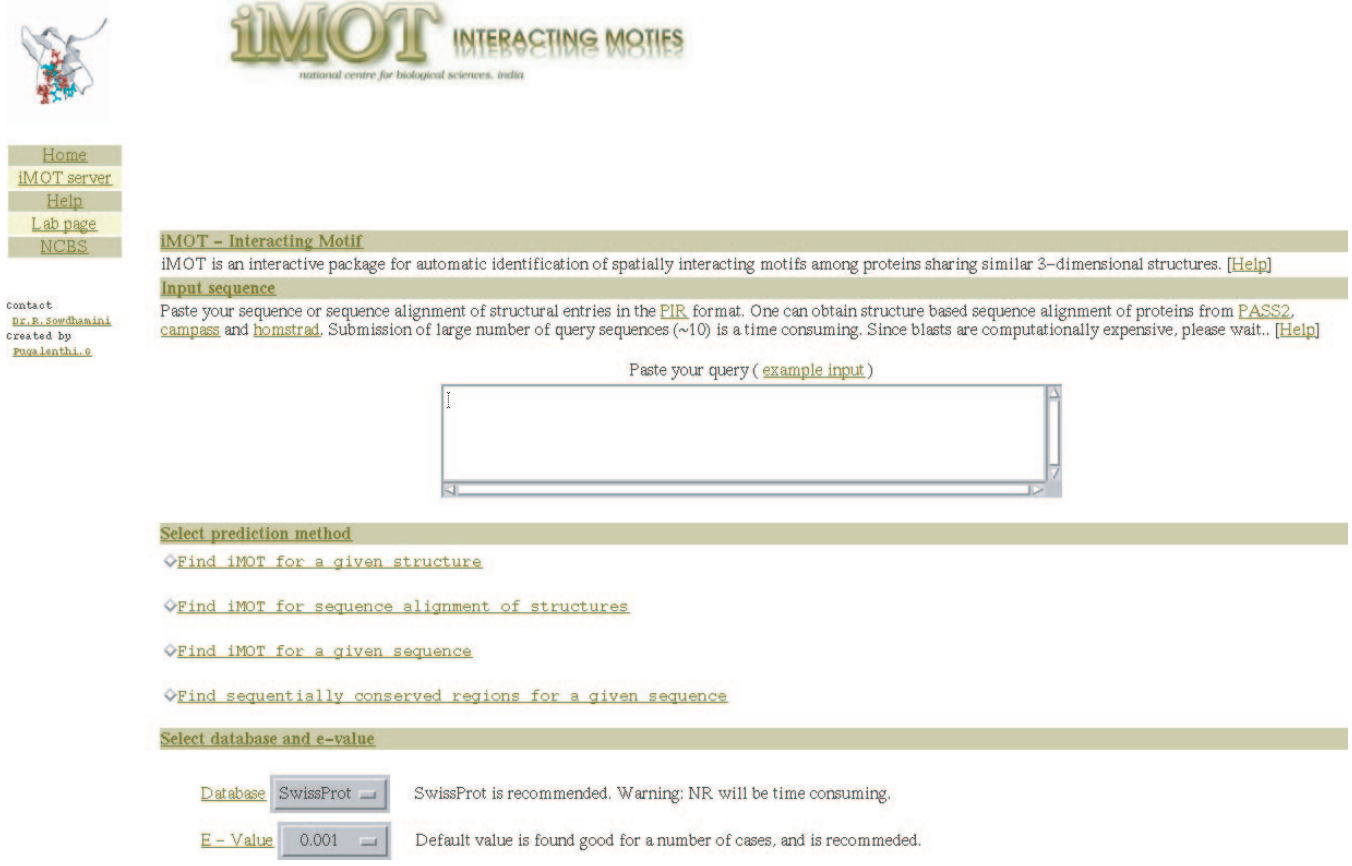

Figure 1. Snapshot of the iMOT server with its various options.

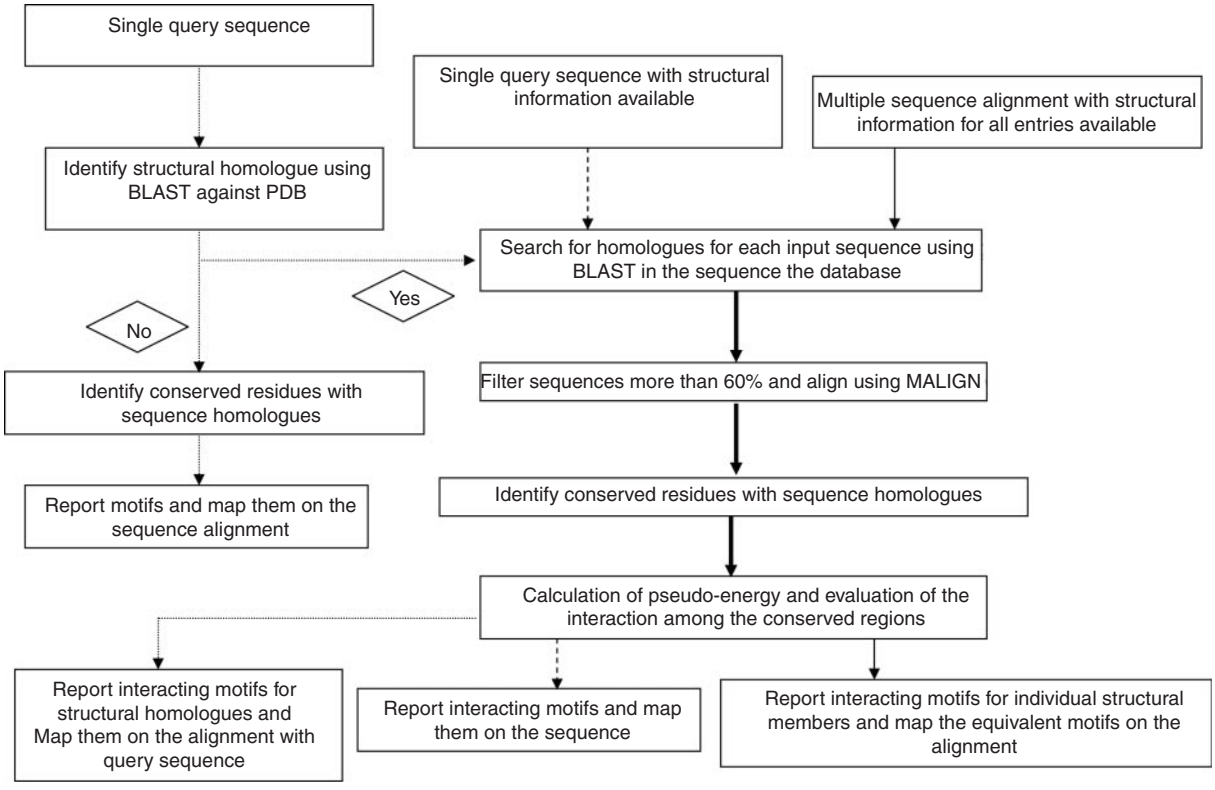

Figure 2. Flowchart representing the steps involved in identification of interacting motifs for a given query. The user may submit a single query sequence, a single query sequence with structural information available or a multiple sequence alignment with structural information for all entries available. After identifying the interacting motifs for a query, the motifs are reported and mapped on the alignment using iMOT. If no structural homologues are present, the conserved regions of the query are reported and mapped on the alignment along with its homologues.

(with sequence identity <40% among themselves) and searched for PROSITE motifs using the ps_scan program (25). The domain definitions corresponded to the SCOP database (6). Functional motifs, as recorded in the PROSITE database, could be identified in 1664 out of 3867 structural domains. In the vast majority (98.2%) of instances, the functional PROSITE motifs formed a subset of the spatially interacting motifs. iMOT provides 52 443 motifs out of which only 3235 are represented in the PROSITE database (see Supplementary Materials).

Motifs of the phosphotyrosine phosphatase IB family: a case study

Tyrosine phosphatases are essential protein modules in signal transduction and are biochemically characterized by the removal of the phosphatase group from phosphorylated tyrosine residue. The catalytic signature of [LIVMF][LIVMF]-H-C-x(2)-G-x(3)-[STC]-[STAGP]-x-[LIVMFY] is conserved across all tyrosine phosphatase family members and has been documented in PROSITE as a functional motif. iMOT suggested 16 motifs when one of the classical tyrosine phosphatases, human PtpIB structure (PDB code 2hnq), was provided as a query (see Supplementary Materials). Three of these motifs, not documented in earlier studies (26) to be either functionally or structurally important, are identified by this approach. Such motifs may contribute further to the structural integrity of the fold.

CONCLUSIONS

Homology among proteins often implies similarities in their structural and functional properties. Structural and functional fingerprints are conserved among homologues. Diverse protein families often lack sequential motifs and are represented through Hidden Markov Models (27) that are sensitive to the detection of distant relationships but are not applicable to ascribing critically important residues to a protein. Interacting motifs provide reliable patterns representing diverse families. Interacting motifs are essential components of proteins that aid retention of function or are structurally specific to the superfamily. They are of potential use in protein engineering and modelling experiments, since they may be crucial features for structural stability and biochemical function, forming an integral part of the protein.

ACKNOWLEDGEMENTS

R.S. is a Senior Research Fellow of the Wellcome Trust, UK. The stay of P.G. in NCBS was supported by Wellcome Trust as part of R.S.'s fellowship. We also thank NCBS (TIFR) for infrastructural support.

REFERENCES

1. Anfinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

2. Doolittle,R.F. (1981) Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149–159.
3. Reddy,B.V., Li,W.W., Shindyalov,I.N. and Bourne,P.E. (2001) Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins. *Proteins*, **42**, 148–163.
4. Sowdhamini,R., Burke,D.F., Huang,J.F., Mizuguchi,K., Nagarajaram,H.A., Srinivasan,N., Steward,R.E. and Blundell,T.L. (1998) CAMPASS: a database of structurally aligned protein superfamilies. *Structure*, **6**, 1087–1094.
5. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
6. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
7. Bairoch,A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.*, **19** (Suppl.), 2241–2245.
8. Henikoff,S. and Henikoff,J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.*, **19**, 6565–6572.
9. Attwood,T.K. and Beck,M.E. (1994) PRINTS: a protein motif fingerprint database *Protein Eng.*, **7**, 841–848.
10. Cardle,L. and Dufton,M.J. (1994) Identification of important functional environs in protein tertiary structures from the analysis of residue variation in 3D application to cytochromes-c and carboxypeptidase-a and carboxypeptidase-b. *Protein Eng.*, **7**, 1423–1431.
11. Bhaduri,A., Ravishankar,R. and Sowdhamini,R. (2004) Conserved spatially interacting motifs of protein superfamilies: application to fold recognition and function annotation of genome data. *Proteins*, **54**, 657–670.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
13. Johnson,M.S. and Overington,J.P. (1993) A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J. Mol. Biol.*, **233**, 716–738.
14. Raschke,T.M., Tsai,J. and Levitt,M. (2001) Quantification of the hydrophobic interaction by simulations of the aggregation of small hydrophobic solutes in water. *Proc. Natl Acad. Sci., USA*, **98**, 5965–5969.
15. Bairoch,A. and Apweiler,R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.*, **24**, 21–25.
16. Bourne,P.E., Addess,K.J., Bluhm WF, Chen,L., Deshpande,N., Feng,Z., Fleri,W., Green,R., Merino-Ott,J.C., Townsend-Merino,W. *et al.* (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res.*, **32**, D223–D225.
17. Johnson,M.S., Overington,J.P. and Blundell,T.L. (1993) Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol.*, **231**, 735–752.
18. Mizuguchi,K., Deane,C.M., Blundell,T.L., Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families *Protein Sci.*, **7**, 2469–2471.
19. Overington,J.P., Zhu,Z.Y., Sali,A., Johnson,M.S., Sowdhamini,R., Louie,G.V. and Blundell,T.L. (1993) Molecular recognition in protein families: a database of aligned three-dimensional structures of related proteins. *Biochem. Soc. Trans.*, **21**, 597–604.
20. Johnson,M.S. and Overington,J.P. (1993) A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J. Mol. Biol.*, **233**, 716–738.
21. Novotny,J., Bruccoleri,R.E. and Saul,F.A. (1989) On the attribution of binding energy in antigen–antibody complexes McPC 603, D1.3 and HyHEL-5. *Biochemistry*, **28**, 4735–4749.
22. Novotny,J., Bruccoleri,R.E., Davis,M. and Sharp,K.A. (1997) Empirical free energy calculations: a blind test and further improvements to the method. *J. Mol. Biol.*, **268**, 401–411.
23. Dimitrov,R.A. and Crichton,R.R. (1997) Self-consistent field approach to protein structure and stability. I: pH dependence of electrostatic contribution. *Proteins*, **27**, 576–596.
24. Lomize,A.L., Reibarkh,M.Y. and Pogozheva,I.D. (2002) Interatomic potentials and solvation parameters from protein engineering data for buried residues. *Protein Sci.*, **11**, 1984–2000.
25. Gattiker,A., Gasteiger,E. and Bairoch,A. (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl. Bioinformatics*, **1**, 107–108.
26. Andersen,J.N., Mortensen,O.H., Peters,G.H., Drake,P.G., Iversen,L.F., Olsen,O.H., Jansen,P.G., Andersen,H.S., Tonks,N.K. and Moller,N.P. (2001) Structural and evolutionary relationships among protein tyrosine phosphatase domains. *Mol. Cell Biol.*, **21**, 7117–7136.
27. Bucher,P. and Bairoch,A. (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 53–61.